DOI:10.11918/202108065

混合跨域神经网络的草图检索算法

李奇真[1],周 圆[2],李 绰[2],彭一南[2],梁先明[1]

(1.中国电子科技集团第十研究所,成都 610036;2.天津大学 电气自动化与信息工程学院,天津 300072)

摘 要: 基于草图的跨域图像检索任务以手绘草图为输入,从彩色图像数据库中检索得到最相似的图像。为了在基于草图的图像检索任务中,更好地融合来自草图和彩色图像的特征,本文提出了用于草图检索任务的混合跨域神经网络,由草图特征提取分支与异构特征融合的彩色图像网络分支组成。该网络提取获得手绘草图、正负样本彩色图像及其边缘轮廓的特征表示,并将彩色图像及其草图近似图(即彩色图像的边缘轮廓)进行特征融合,作为彩色图像特征,弥补了手绘草图与彩色图像直接匹配的跨域差距。通过对网络模型的参数与网络结构等方面探索,进一步优化草图检索算法。在 Flickr15K 草图检索数据集上的实验结果表明,本文提出的方法优于当前其他先进的草图检索算法,在检索平均精确度这个客观指标上达到了 0.584 8,相比其他方法中指标最优的值提升了 0.052 2。

关键词: 草图检索;跨模态;神经网络;图像检索

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 0367-6234(2022)05-0064-10

Hybrid cross-domain joint network for sketch-based image retrieval

LI Qizhen[1], ZHOU Yuan[2], LI Chuo[2], PENG Yinan[2], LIANG Xianming[1]

(1. Tenth Institute of China Electronics Technology Group Corporation, Chengdu 610036, China;
2. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Sketch-based cross-domain image retrieval (SBIR) uses a sketch as query to retrieve the most similar image from the color image database. In this study, in order to better fuse the features from sketch and color image, a hybrid cross-domain joint network for sketch-based image retrieval was proposed, consisting of a sketch feature extraction branch and a color image heterogeneous feature fusion network branch. The network extracts the feature representations of sketch, positive and negative color image, and corresponding edge outline, and fuses the features of the color image and its sketch approximation (the edge outline of the color image) as the color image feature, which bridges the cross-domain gap between sketches and images. The network model parameters and network structure were further explored to optimize the purposed algorithm. Experiment on Flickr15K dataset shows that the proposed method performed better than other advanced image retrieval methods. The mean average retrieval accuracy of the proposed method was 0.584 8, which was 0.052 2 higher than the optimal value in other methods.

Keywords: sketch retrieval; cross-modal; neural networks; image retrieval

随着互联网与图像资源的快速发展,人们对图像语义的表达、理解和查询更加困难[1]。如何从海量图片中快速找到目标图片成为检索任务的重要问题(例如近似图像检索[2-3]、草图检索[4-5])。图像检索最初的研究方向为基于文本的图像检索(text-based image retrieval,TBIR)[6-7],但是由于文本描述的不全面性与图片标注的复杂性,图像检索向基于内容的图像检索(content-based image retrieval,CBIR)[8-9]等方向不断发展。

草图作为一种方便表达的方式,因其广泛的应用场景,例如草图上色[10]、草图语义分割[11]、草图辅助真实图像生成[12],受到研究者的关注。基于草图的跨域图像检索旨在根据输入手绘草图与数据库中彩色图像的相似性度量返回彩色图像排序结果。该检索通过手绘草图作为查询信息完成图像检索,可应用在基于文本的图像检索中文本不准确时,或基于内容的图像检索中,彩色图像受到场景和拍摄条件限制难以获得时。基于草图的图像检索(sketch based image retrieval,SBIR)直观、高效、便捷,更符合用户的表达习惯,自提出以来就受到持续关注[13],在图片检索和其他跨域检索方面都发挥着重要作用[14-15]。

基于传统方法的草图检索首先从彩色图像中提取边缘轮廓图代替彩色图像作为待查询数据库,然后使用传统方法(如方向梯度直方图(HOG)[16]、面

收稿日期: 2021-08-16
基金项目: 国家重点研发计划(2020YFC1523204);
国家自然科学基金(62171320,U2006211)
作者简介: 李奇真(1989—),男,博士,工程师
通信作者: 周 圆,zhouyuan@tju.edu.cn;
梁先明,liangxianming21@163.com

向草图特征的方向梯度直方图(SHOG)[17]、基于梯度场的方向梯度直方图(GF-HOG)[18-20]、边缘局部直方图[21]、边缘局部软直方图(SHELO)[22-23]和边缘感知[24])从草图和彩色图像(或彩色图像轮廓图)中提取浅层特征，并将提取的特征经过视觉词袋(bag of visual words，BoVW)[25]等方法进一步整合特征。最后通过查询草图特征和数据库中图像特征间的相似性度量实现草图检索。然而，传统特征的稀疏草图和彩色图像之间的跨域差距没有充分解决，手绘草图的抽象性和绘画风格的多变性使检索效果不佳。

深度学习自发展以来解决了许多视觉问题，如图像显著性检测[26-27]、目标跟踪[28]，基于深度学习的草图检索方法也能取得比原来更好的效果。2009 年，Eitz 等[29]构建了 TU-Berlin 手绘草图数据集；Yu 等[30]提出首个针对草图分类的卷积神经网络 Sketch-a-Net(SaN)，在 AlexNet[31]基础上采用更大的卷积核与池化步长，适应草图线条的稀疏性，其分类准确率明显高于传统方法；2016 年，研究者通过向孪生网络框架引入 SaN，将草图与彩色图像轮廓图映射到同一特征空间以进行相似图像聚类[32-33]；2017 年，Bui 等[34]以草图作为锚点，彩色图像轮廓图作为正负样本，并以 SaN 为基础网络组成三元组网络实现草图检索[34-36]；结合自步学习[37]与课程学习[38]，Xu 等[39]提出一种新型草图检索方法。

基于深度学习的 SBIR 方法目前存在的主要问题有：1)草图是抽象的描述，与彩色图像有本质的区别，将草图与彩色图像映射到同一目标特征空间是十分困难的；2)当彩色图像的背景相对复杂时，其轮廓图将含有较多背景轮廓，不能真实反应彩色图像的内容信息，此时彩色图像轮廓图与草图将难以匹配。

因此本文提出一种新型混合跨域神经网络，将草图域、草图近似图域、彩色图像域的图像共同输入网络，挖掘 3 个域间的共有特征。通过增加草图近似图域作为中间表示，缓解查询草图和被检索彩色图像之间的特征跨域问题。通过研究特征提取网络结构和数据扩充算法，设计了一种基于深度学习的 SBIR 方法。

1　混合跨域神经网络的结构

本文提出了一种由三元组损失函数作为监督的混合跨域网络结构，见图 1。该网络由草图分支、正样本彩色图像分支与负样本彩色图像分支组成。草图分支由 Sketch-a-Net 网络的特征提取模块构成，即图 1 中间位置的分支，用于提取草图域特征。正负样本彩色图像分支完全相同，均由异构融合网络构成，如图 1 上下位置的两个大分支以及图 2 所示，分别处理彩色图像域和草图近似图域的图像。下面对混合跨域神经网络的内部结构进行详细介绍。

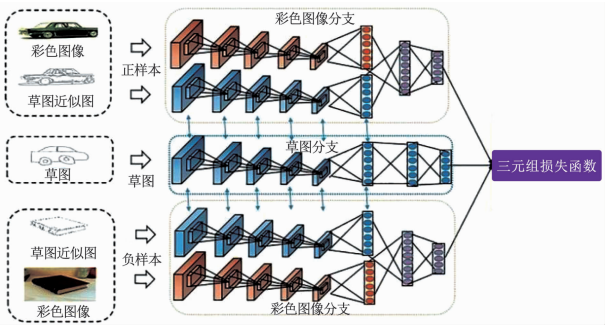

图 1　混合跨域神经网络结构

Fig. 1　Structure of hybrid cross-domain network

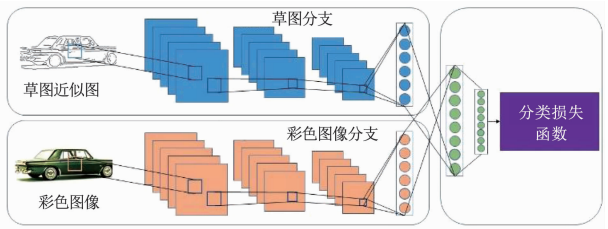

图 2　异构融合网络结构

Fig. 2　Structure of heterogeneous fusion network

1.1　异构特征融合网络

彩色图像特征提取网络分支由两支并行的卷积神经网络与最后共用的全连接层组成，见图 2。异构特征融合网络的上半部分为一支草图近似图输入网络，该网络包含 5 个卷积层和 2 个全连接层，分别对应 SaN 的卷积层和 AlexNet 全连接层。异构特征融合网络的下半部分为一支彩色图像输入网络，同样由 5 个卷积层和 2 个全连接层组成，与 AlexNet 具有相似的网络设置。以上两支基础神经网络具体细节见表 1。

表 1　异构融合网络两个分支的具体结构

Tab. 1　Detailed structures of two branches in heterogeneous fusion network

层	网络			
	彩色图像		草图/草图近似图	
	卷积核大小	输出通道数	卷积核大小	输出通道数
Conv1	16×16	64	11×11	96
MaxPool	3×3		3×3	
Conv2	6×6	128	5×5	256
MaxPool	3×3		3×3	
Conv3	3×3	256	3×3	384
Conv4	3×3	256	3×3	384
Conv5	3×3	256	3×3	256
MaxPool	3×3		3×3	
FC6	7×7	4 096	6×6	4 096
FC7	1×1	4 096	1×1	4 096

受多模态特征学习[40-41]的启发,为了获得彩色图像与草图近似图的公共特征表示,在全连接层对草图近似图特征与彩色图像特征进行特征融合,见图2。然后使用融合后的异构融合网络对草图近似图与彩色图像进行特征提取,将提取到的跨域共享特征作为彩色图像特征。因此得到的共享特征能够更充分地代表彩色图像,以弥补草图与彩色图像之间的跨域差异。

1.2 草图网络

由于手绘草图与草图近似图均由相似的线条信息组成,因此本文使用具有相同配置的网络结构对草图与草图近似图特征进行提取,见表1。然而,相对于草图近似图较为规则的轮廓信息,手绘草图的线条信息更为抽象。因此,本文通过对草图与草图近似图网络进行部分参数共享网络设置,解决手绘草图与草图近似图之间的差异问题。通过对草图与草图近似图进行相同网络与参数半共享设置,使得草图与草图近似图能够较好地保留原本的图像特征,同时增加草图与草图近似图之间的相似性,从而充分发挥草图近似图作为桥接草图和彩色图像跨域信息的重要作用。

1.3 排序损失函数

草图检索的最终目标是得到手绘草图与彩色图像的相似性度量模型,该模型首先通过神经网络对草图与彩色图像进行特征提取,然后通过相似性度量在特征空间对特征进行匹配排序,并且使得与查询草图相似的彩色图像排序靠前,而不相似的图片排序靠后。在此过程中,本文使用欧几里得距离对草图 a 与彩色图像 p 进行相似性度量,相似性度量距离公式为

$$D(a,p) = \| f_\theta(a) - g_\theta(p) \|_2^2 \tag{1}$$

式中:$f_\theta(\cdot)$和 $g_\theta(\cdot)$分别为提取并映射到特征公共空间的草图与彩色图像特征提取网络近似函数,$D(\cdot,\cdot)$为草图特征 $f_\theta(a)$和 $g_\theta(p)$之间相似性距离度量。

为了得到草图与彩色图像更好的特征表示 $f_\theta(a)$和 $g_\theta(p)$,需要对特征提取网络进行训练。本算法使用跨域混合网络结构结合三元组损失函数对网络的输出特征和$\{f_\theta(a), g_\theta(p^{I+}, p^{S+}), g_\theta(n^{I-}, n^{S-})\}$进行监督,对特征提取网络进行训练,见图1。其中 $f_\theta(a)$为草图特征提取网络分支的输出特征,$g_\theta(p^{I+}, p^{S+})$为异构特征融合网络正样本彩色图像的输出特征,$g_\theta(n^{I-}, n^{S-})$则为异构特征融合网络负样本彩色图像的输出特征。实验的最终目标是在公共特征映射空间中,使得正样本彩色图像特征与查询草图特征更相近,而负样本彩色图像特征与草图特征距离更远,其表达式为

$$D(f_\theta(a), g_\theta(p^{I+}, p^{S+})) < D(f_\theta(a), g_\theta(n^{I-}, n^{S-})) \tag{2}$$

对应的三元组损失函数为

$$L = \max(0, m + \| f_\theta(a) - g_\theta(p^{I+}, p^{S+}) \|_2^2 - \| f_\theta(a) - g_\theta(n^{I-}, n^{S-}) \|_2^2) \tag{3}$$

式中 m 为度量正样本彩色图像到草图之间的距离 $D(f_\theta(a), g_\theta(p^{I+}, p^{S+}))$与负样本彩色图像到草图之间距离 $D(f_\theta(a), g_\theta(n^{I-}, n^{S-}))$差的阈值。即输出特征应在满足式(2)的基础上,同时使得 $D(f_\theta(a), g_\theta(p^{I+}, p^{S+}))$与 $D(f_\theta(a), g_\theta(n^{I-}, n^{S-}))$之差大于一个阈值。当满足以上条件时,在公共特征空间内正负样本将会被正确排序,且有一定的区分度。否则,三元组损失函数将产生一个0~1之间的损失并对网络模型进行惩罚训练[43]。总的损失函数定义如下:

$$\min \sum_i L_i + \lambda W(\theta) \tag{4}$$

式中:i 为训练时输入的三元组图像对的序列号,θ为深度神经网络的参数,$W(\theta)$为防止网络模型过拟合的最小平方误差损失正则项,λ 为正则化参数。最小化损失函数在缩小草图与彩色图像正样本之间的距离同时增大草图与彩色图像负样本之间的距离。随着网络模型的训练,最终将学习到一个较好的排序模型并用于最终的草图与彩色图像的特征匹配。

2 混合跨域神经网络的训练测试过程

2.1 数据集获取与增强

已有实验[42-43]表明:在深度学习算法中,训练数据的数量和质量是影响算法效果的重要因素之一。因此在基于草图的图像检索中,也存在着类似的问题。目前,基于草图的图像检索受到手绘草图数据少且难以获得等因素的影响,网络的训练性能并不理想。此外,草图和彩色图像之间存在着两种不同的视觉表现形式,草图和彩色图像之间的差异使得基于草图的跨域图像检索成为了一项非常具有挑战性的任务。因此,本文提出通过生成彩色图像边缘轮廓作为草图近似图的方法以解决以上问题。草图近似图的生成操作有两方面优势:1)扩增草图数据集,使用草图近似图近似表示草图以扩充草图数据集,为草图检索训练过程中出现的草图数据集不足的问题提供了一个新思路;2)缩小域间差异,本文使用草图近似图与彩色图像进行特征融合,融

合后的特征作为彩色图像特征；该特征同时具有手绘草图的线条信息与彩色图像的细节信息，因此能够减小草图和彩色图像之间直接的跨域差。

在此之前基于草图的图像检索算法[16-18,20,44]中主要使用 Canny 边缘检测算法对彩色图像进行轮廓提取，但是 Canny 算法生成的边缘轮廓图中存在较多的冗余信息。因此，本文使用目前较为先进的边缘检测算法（generalized boundary detector，Gb）[45]对彩色图像进行边缘轮廓提取，并生成草图近似图。Gb 算法有效地组合了输入图像的低级和高级语义特征来完成彩色图像的边缘轮廓提取，但是在使用 Gb 算法进行轮廓提取之后，提取的边缘轮廓中仍然包含较多的细节。因此，本文在使用 Gb 算法提取边缘图像之后，通过设置阈值对图像中较弱边缘像素进行移除，留下强边缘像素。不同阈值的选取会产生不同抽象程度的草图近似图，通过设定不同的阈值，可以获得不同边缘保留度的草图近似以扩增草图数据集。图 3 分别为网络训练过程中所需的手绘草图、彩色图像以及通过预处理后产生的草图近似图。此外，草图、彩色图像及由其生成的草图近似也会通过随机旋转水平翻转的方式进行数据增强。这些数据增强也可以补偿有限的可用数据以及训练过程中出现的过拟合问题。

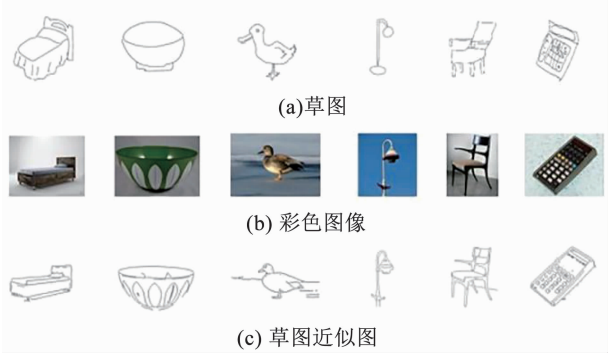

图 3　混合网络输入示意

Fig. 3　Illustration of hybrid network inputs

2.2　训练过程

受到网络结构和训练数据的复杂性等限制，本实验使用多步训练的方式对所提出的跨域混合网络进行训练。训练过程如下：

步骤 1　基础网络预训练。首先使用 softmax 分类损失函数对手绘草图/草图近似图输入网络（图 4（a）左侧分支）和彩色图像输入网络（图 4（a）右侧分支）两个基础网络进行分类预训练。手绘草图/草图近似图输入网络将在 SaN 模型上进行微调，彩色图像输入网络将在 AlexNet 训练模型上进行微调。

步骤 2　异构融合网络预训练。将步骤 1 中训练好的手绘草图/草图近似图输入网络（图 4（a）左侧分支）和彩色图像输入网络（图 4（a）右侧分支）前 7 层作为异构融合网络（图 4（b））的基础网络，对异构融合网络进行训练。为了获得较好的彩色图像特征，在异构融合网络最后一层加入分类损失函数对高层语义进行分类预训练。预训练过程中异构融合网络的输入分别为彩色图像及其对应的轮廓图。

步骤 3　混合网络训练。步骤 1、2 主要对草图分支（草图输入网络）与彩色图像分支（异构融合网络）进行了分类判别训练。为了进一步得到草图与彩色图像之间更好的跨域排序模型，步骤 3 将对草图分支与彩色图像分支组成的混合网络结构进行训练，见图 4（c）。首先移除草图分支（图 4（a）左侧分支）与彩色图像分支（图 4（b）异构融合网络）中的分类层；然后将两支网络作为混合网络的基础网络，结合三元组损失函数对混合网络进行训练。

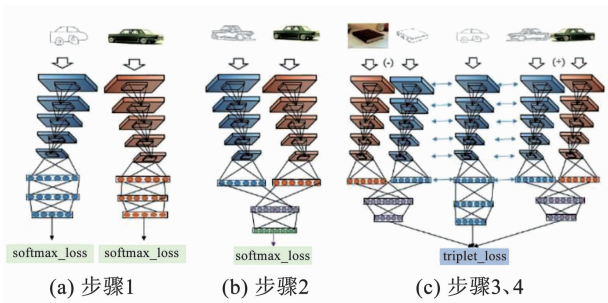

图 4　具体训练过程

Fig. 4　Detailed training procedure

步骤 4　混合网络微调。受到训练数据的限制，在条件允许的情况下使用尽可能多的数据集（如：草图-草图近似图-彩色图像或者草图近似图-草图近似图-彩色图像）对混合网络（图 4（c））进行微调。

以上为基于混合跨域神经网络的草图检索算法的训练过程，之后可以将训练好的网络模型用于测试阶段特征的提取与相似性匹配。

2.3　测试过程

如上所述，草图检索的最终目标是得到手绘草图与彩色图像的相似性度量模型。因此，在完成跨域混合网络模型的训练之后（图 5），本节使用训练好的草图分支（草图输入网络）与彩色图像网络分支（异构融合网络）分别对草图与彩色图像进行特征提取；然后对提取到草图特征与跨域融合的彩色图像特征进行相似性度量匹配；最后，根据相似度得分进行排序，并返回排序结果。

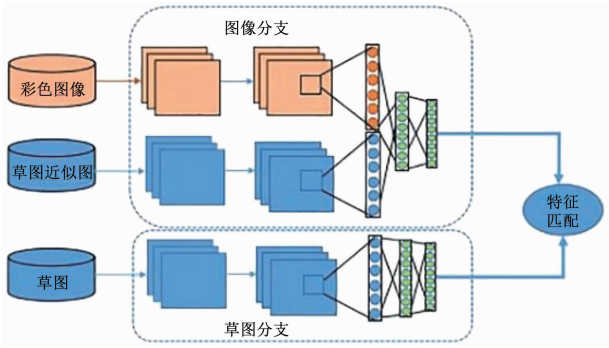

图 5 测试过程网络结构

Fig. 5 Illustration of test procedure

3 实验及结果

本节首先介绍实验所需的训练和测试数据集，然后描述了数据预处理和实现细节，接着将本方法与其他对比方法在广泛使用的 SBIR 数据集上进行了客观指标对比，其次对本文方法检索的结果进行主观展示，最后对所提网络结构进行消融研究。

3.1 数据集

TU-Berlin[46]是第一个大型手绘草图数据集。该数据集共 20 000 张草图，主要分为 250 类，每类 80 张，分别由 1 350 位非专业人员绘制而成。

Flickr25K[34]包含 25 000 张彩色图像的数据集。为了与 TU-Berlin 组成草图-彩色图像输入对以进行草图检索网络训练，Bui 等[34]从 Flickr、API、Google 和 Bing 等平台搜集到与 TU-Berlin 具有相同类别数的彩色图像。因此，该数据集由 250 类彩色图像组成，每类 100 张。

Flickr15K[18]包含测试草图 330 张与彩色图像 14 460 张。该数据集分为 33 类，每类 10 张，由非专业人员绘制而成，共 330 张草图用于最终的测试输入。该数据集还有 14 460 张彩色图像，根据形状相似性被分为对应的 33 类，每类数据不均衡。

与文献[29,32－34]数据处理相似，本文将彩色图像平均分成两部分，其中一部分用于测试时被检索的彩色图像数据库数据，另外一部分用于训练过程步骤 4 网络微调，而草图全部用于测试阶段。由于微调网络需要有手绘草图，因此本文使用数据增强方法，对彩色图像进行轮廓提取，将提取到的草图近似图作为草图数据集来解决缺少草图训练集的问题。

综上所述，在网络训练的步骤 1 中，实验使用 TU-Berlin 中 20 000 张草图与 Flickr25K 中 25 000 张彩色图像分别对草图/草图近似图输入网络与彩色图像输入网络进行分类预训练。步骤 2 中使用 Flickr25K 中 25 000 张彩色图像及其对应的草图近似图对异构融合网络进行分类预训练。步骤 3 中，草图输入部分使用 TU-Berlin，正样本的输入为 Flickr25K 中与草图输入具有相同类别的彩色图像及其对应的草图近似图，负样本为 Flickr25K 中与草图输入具有不同类别的彩色图像及其对应的草图近似图。最后，在步骤 4 中只使用 Flickr15K 中彩色图像训练数据及其提取到的草图近似图部分作为特征融合网络的彩色图像正负样本分支的输入，草图近似图作为草图网络的输入。

3.2 实验设置

3.2.1 数据集预处理与增强

实验的预处理过程使用带有不同阈值的 Gb 算法获得草图近似图，并将提取的草图近似图分别用于草图网络与异构融合网络中草图近似图网络的输入。为了扩增草图数据，实验设置使用 5 种阈值：0.5、0.6、0.7、0.8 和 0.9 进行边缘轮廓提取，以获得 5 种不同细节保留度的草图近似图。最终使用 7 330 × 5 = 36 650 张草图近似图来扩充草图数据集进行训练。同时，实验使用阈值为 0.9 获得的草图近似图作为异构融合网络中草图近似图网络分支的输入。

在实验过程中，为了获得网络输入所需大小的输入图像同时实现数据增强。实验首先对草图、彩色图像及其对应草图近似图尺寸统一调整为 256 × 256 的大小。然后，将草图及草图近似图随机裁剪为 225 × 225 的大小以适应草图网络的输入，彩色图像则被随机裁剪为 224 × 224 的大小以适应彩色图像网络的输入。

3.2.2 实验参数细节

本文使用 caffe 深度学习框架搭建神经网络模型，使用随机梯度下降(stochastic gradient descent, SGD)的方式进行网络训练，训练硬件环境为 NVIDIA TITAN Xp GPU。式(4)中的正则项参数 $\lambda = 0.000\ 5$。网络最初的训练速率为0.005，微调的训练速率为 0.000 1。实验使用0.000 5 的权重衰减与 0.9 的动量进行网络参数调整，并使用 MATLAB 2015b 实现草图近似图的提取。在测试阶段，本文使用 mAP (mean average precision)作为客观评价指标。mAP 是一种检索任务通用的客观指标，它并不是直接使用“检索成功的数目除以检索结果总数”来计算。在检索任务中，检索成功的例子出现在检索结果的不同排序位置上，mAP 对于结果也有着不同的影响。在后续段落，mAP 指标以 I_{mAP} 表示，其表达式为

$$I_{\mathrm{mAP}} = \frac{\sum_{q=1}^{Q} I_{\mathrm{AP}}(q)}{Q} \tag{5}$$

$$I_{AP} = \frac{1}{n} \times \sum_{r=1}^{n} \frac{r}{p(r)} \tag{6}$$

式中：I_{AP}为第 r 个检索结果匹配分数；n 为一次检索返回的多个结果中正确检索的结果个数；$p(r)$为检索结果中，从前往后的第 r 个检索成功的信息所在位置。例如：在返回的 6 个检索结果中，有 3 个与查询信息具有相同标签（即检索成功，$n=3$），这些检索成功结果的所在位置为 1、3、6，则 $I_{AP} = \frac{1}{3} \times (\frac{1}{1} + \frac{2}{3} + \frac{3}{6})$。

3.3　对比实验及结果分析

本节将所提出的混合跨域网络的检索性能与基于 Flickr15K[18] 数据集进行训练的其他 SBIR 算法的检索性能进行比较。对比算法包括：

基于传统特征的 SBIR 方法：草图直方图（sketch histogram of oriented gradients，SHOG）[17]、基于梯度场的梯度直方图（gradient field histogram of oriented gradients，GF－HOG）[18]、感知边缘（perceptual edge）[24] 和基于学习的关键形状（learned key shapes，LKS）[23]。

基于深度学习特征的 SBIR 方法：草图网络（sketch-a-net，SaN）[30]、孪生网络（Siamese network）[32]、带有 LKS 特征的跨步学习神经网络（cross-paced representation learning convolutional neural network with learned keyshapes，CPRLCNN＋LKS）[39]、非对称性特征图（asymmetric feature map，AFM）[47]、带有重排序的对象层级的草图网络描述子（object level sketch-a-net descriptor with reranking）[48]、基于四元组的检索网络（quadruplet multi-task，Quadruplet MT）[49]、查询自适应重排序卷积神经网络（query-adaptive re-ranking CNN）[44]，以及基于三元组网络的方法（triplet network）[34－36]。

实验结果见表 2 及图 6。由表 2 可以看出，本文所提出的跨域混合网络结构相比其他方法有明显的效果提升。在深度学习方法中，相对于使用相同深度基础网络结构的算法[36]，I_{mAP}指标提升了 0.17，而对于使用更深基础网络结构的算法，本文算法在 I_{mAP}指标上仍有 0.05 的效果提升。同时，与最好的传统草图检索方法 LKS 相比，I_{mAP}指标提升了 1 倍多。图 6 是本文草图检索方法的部分结果，从检索结果中挑选一些具有代表性的结果进行展示。其中，第 1 列为输入的查询草图，后面几列为查询草图对应的彩色图像检索结果，框出部分为错误的检索结果。从图 6 可以看出，即使检索得到的结果标签错误，其仍与查询草图在形状上十分相似。由实验结果可以证明本文提出的跨域混合网络具有较好的检索效果。

表 2　不同方法在 I_{mAP}指标上的对比

Tab. 2　Comparison of different methods on I_{mAP}

方法名称	I_{mAP}
SHOG[17]	0.109 3
GF-HOG[18]	0.122 2
Perceptual Edge[24]	0.183 7
Learned key shapes（LKS）[23]	0.245 0
SaN[30]	0.173 0
Siamese CNN[32]	0.195 4
Triplet（sketch-edge map）[34]	0.244 5
CPRLCNN＋LKS[39]	0.273 4
Asymmetric feature map（AFM）[47]	0.304 0
Object level Sketch-a-Net descriptor＋Reranking[48]	0.319 0
Quadruplet MT[49]	0.321 6
Query-adaptive re-ranking CNN[44]	0.323 0
Triplet（sketch-image）[35]	0.359 1
Triplet（sketch-image with fine-tuned final model）[36]	0.374 1
Triplet（partial sharing convnet）[50]	0.532 6
本文方法	0.544 8
本文方法（带有查询扩展）	0.584 8

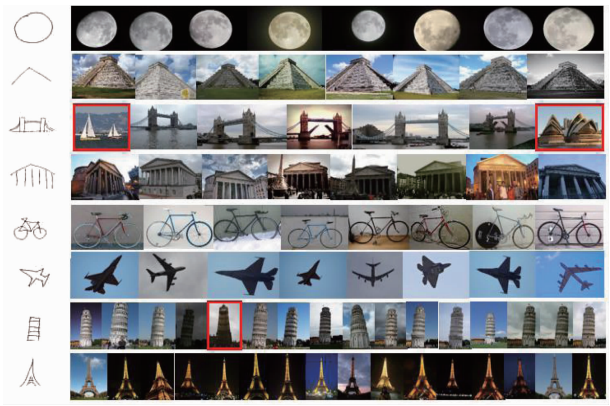

图 6　Flickr15K 数据集上草图检索排序结果示例（框出部分为检索错误的图片）

Fig. 6　Examples of ranking results of sketch retrieval on Flickr15K dataset（The images with boxes are wrongly retrieved.）

3.4　对网络结构的探索

考虑到网络模型的复杂性，为了得到一个更优的检索结果，从以下几个方面对网络模型进行探索与分析：1）参数共享；2）异构融合网络结构；3）特征降维；4）扩展查询；5）三元组模型对比。

3.4.1　参数共享模式

由于手绘草图与草图近似图的域间差异小，本文使用相同的网络配置对手绘草图与草图近似图进行特征提取。但是手绘草图由抽象的弯曲线条组成，而草图近似图则较为规则，针对这一差异，提出使用半参数共享的方式对网络进行训练。对混合网络中草图与草图近似图网络之间不同的参数共享结构进行对比，探讨网络参数的最佳配置见图 7。由图 7（a）～（i），草图与草图近似图网络参数共享结

构设置依次为:只有 1 层参数共享,1、2 层参数共享设置……直到两支网络完全参数共享,然后从第 1 层开始依次减少共享层直到没有参数共享,即两支网络完全独立。因此,本节探索了 12 种网络参数共享设置,并进行了实验对比,实验结果见图 8。由图 8 可知,混合网络结构中草图与草图近似图网络之间使用 1 ~5 层参数共享设置能够得到 I_{mAP} 值最高 0.024 9 的提升。因此,在训练草图检索的混合网络结构时,设置手绘草图与草图近似图输入网络之间 1 ~5 层卷积层参数共享,而全连接层参数独立。

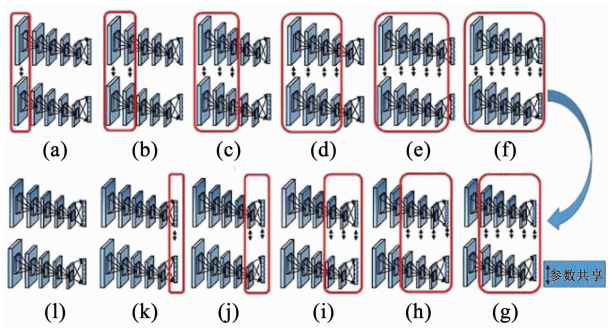

图 7 不同共享层网络结构

Fig. 7 Network structure of different shared layers

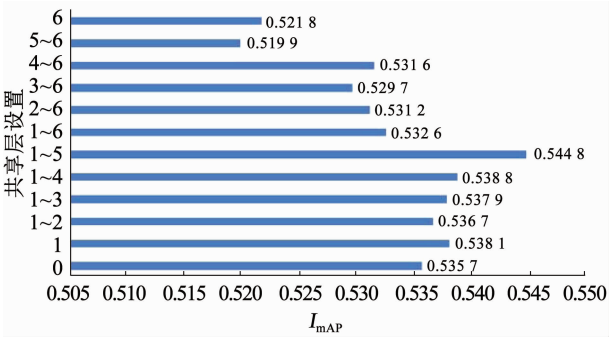

图 8 使用不同层参数共享得到的 I_{mAP}

Fig. 8 I_{mAP} values for different sharing layers

3.4.2 异构融合结构

受到多模态特征学习的启发,本文将彩色图像与草图近似图网络在全连接层进行特征融合,融合后的异构融合网络作为彩色图像分支。由于不同的特征融合方式的会影响彩色图像的特征表达[51],因此本节将探索不同的异构融合网络结构设置,以找到更优的特征提取方式。由图 9(a)可知,网络在第 6 层之后对草图近似图与彩色图像进行特征融合,然后在融合后的下一层进行特征降维。而图 9(b)中,网络在第 7 层后同时进行特征融合与降维操作。两种特征融合网络的区别为:图 9(a)中网络结构尽可能早地将草图近似图与彩色图像的特征进行跨域特征融合,并对跨域融合特征进行训练;而图 9(b)的结构尽可能保持原有域内各自独特的信息,然后再进行特征融合。两种特征融合网络结构均在其他实验设置相同的情况下进行对比分析,实验结果见表 3。

实验结果表明,图 9(a)的网络融合方式相对于图 9(b),I_{mAP} 值有 0.05 的提升。因此,将图 9(a)特征融合方式设置为本文最终训练与测试中异构融合网络的特征融合方式。

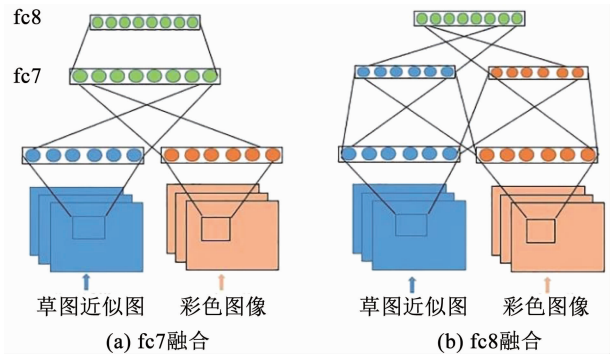

图 9 异构融合网络中两种不同的特征融合方式

Fig. 9 Two different feature fusion methods in heterogeneous fusion network

表 3 不同特征融合网络结构的 I_{mAP} 值

Tab. 3 I_{mAP} values for different feature fusion networks

指标	特征融合	
	fc7 层融合	fc8 层融合
I_{mAP}	0.544 8	0.493 6

3.4.3 特征降维设计

特征降维主要是在特征提取网络后插入一个具有较小节点数的全连接层作为特征输出,以精简特征表达,提高空间占用率并降低计算成本。但是过多的降低特征维数可能会丢失图像的细节信息,因此,本节通过实验探索网络的最佳降维数量。实验结果表明,I_{mAP} 值会随着降维层节点数的增多而增加,而检索时间也会随着节点数量的增加而上升。在计算机设置为 3.40 GHz Intel E3,使用单个 GPU 环境配置下,随着降维数从 64 到 512 的过程中,单张草图检索的时间由 5 ms 上升到 52 ms。考虑到检索时间与检索精度两者的影响,最终选择 128 维作为本实验网络模型输出(I_{mAP} 为 0.544 8,检索时间为 14 ms)。

3.4.4 扩展查询

在进行草图与彩色图像的相似性度量时,本文使用欧几里得距离作为草图与彩色图像的相似性度量标准。为了进一步优化检索结果,降低简单的相似性度量方式带来的匹配误差风险,本文使用拓展查询(query expansion,QE)[52]的方式优化检索排序结果。拓展查询是指计算检索返回的相似度最高前 K 个结果的特征均值作为当前检索的平均特征,以

平均特征作为输入图像特征重新进行检索与排序。在实验中，本节设置最终检索结果前 9 个图像的特征均值作为二次查询的输入特征。图 10 为进行拓展查询前后的排序结果。图中每两列为一组，左边一列为第一次查询结果，右边为拓展查询后的重排序结果，第 1 行均为输入的查询草图，后面几行均为检索输出的排序结果，框出部分为查询错误结果。通过实验结果可以看出，在进行拓展查询调整后，检索排序的前几个结果中将不再有检索错误的图片。通过表 3 可以看出，经过拓展查询操作后检索精度 I_{mAP}值提升了 0.04。

图 10　拓展查询重排序结果示例（框出为检索错误的结果，在进行拓展查询重排序后，检索结果正确）

Fig. 10　Examples of re-ranking results of query expansion (The images with boxes are wrongly retrieved. After the re-ranking process, the retrieval results are corrected.)

3.4.5　与基于三元组模型方法的对比

现存的草图检索方法主要使用草图直接与彩色图像或彩色图像的轮廓图进行匹配。然而，草图与轮廓图的直接匹配效果很大程度上取决于轮廓提取算法的优劣。当图片背景过于复杂或轮廓提取算法不佳时，轮廓图将不能很好地代表彩色图像与草图进行特征匹配。而草图与彩色图像跨域差导致的特征空间分布不一致的现象，使得草图与彩色图像直接进行特征匹配的效果并不理想。因此，本文提出一个新颖的跨域混合网络实现基于草图的图像检索任务。该网络由一个草图分支与两个异构融合网络构成的彩色图像分支组成。相比于之前单一的特征提取网络，使用异构融合网络作为彩色图像的分支能够更好地获取彩色图像的细节信息和草图近似图的线条信息。此时，异构融合网络提取到的融合特征将能够更好地表示彩色图像特征以缩小彩色图像与草图的跨域差距。

如表 4 所示，相比于只用草图与轮廓图或者只用草图与彩色图像组成的三元组网络，本文提出的由 5 支网络组成的混合网络结构在基于草图的图像检索效果上有了明显的提升。由于在基于深度学习的方法中，特征提取效果会随着网络的加深加宽而不断上升。因此，为了公平起见，本实验的三元组基础网络均使用相同深度的神经网络对特征进行提取。

表 4　不同网络结构的 I_{mAP} 值

Tab. 4　I_{mAP} values for different network structures

方法	I_{mAP}
本文方法（草图-草图近似图-彩色图像）	0.544 8
三元组（草图-彩色图像）[36]	0.374 1
三元组（草图-草图近似图）[34]	0.244 5

4　结　论

本文提出了一个新颖的混合网络结构，用于基于草图的跨域图像检索。该混合网络结构由草图网络分支与两个异构融合的彩色图像分支组成。其中，异构融合网络将彩色图像与草图近似图在全连接层进行特征融合作为彩色图像特征。融合后的图像特征将具有彩色图像的细节信息与草图近似图的轮廓信息，从而弥补草图与彩色图像的跨模态差距，最后使用三元组排序损失函数作为监督对混合网络进行训练。此外，通过对参数与网络结构等模型分析与探索进一步优化了网络性能。在 Flickr15K 数据集上的实验结果表明，本文提出的方法优于当前其他先进的草图检索算法，在检索平均精确度这个客观指标上达到了 0.584 8，相比其他方法中指标最优的值提升了 0.052 2。在本文提出的网络的基础上，如果能将诸如文本等来自更多模态的信息引入网络中，并使用多模态特征融合方法进行特征融合，可以在原有检索效果的基础上有一定的性能提升。

参考文献

[1] ZHOU Yuan, WANG Ruolin, LI Hongru, et al. Temporal action localization using long short-term dependency [J]. IEEE Transactions on Multimedia, 2020: 1. DOI: 10.1109/TMM.2020.3042077

[2] XIE Hongtao, GAO Ke, ZHANG Yongdong, et al. Efficient feature detection and effective post-verification for large scale near-duplicate image search[J]. IEEE Transactions on Multimedia, 2011, 13(6): 1319. DOI: 10.1109/TMM.2011.2167224

[3] ZHOU Rong, CHEN Liuli, ZHANG Liqing. Sketch-based image retrieval on a large scale database[C]//Proceedings of the 20th ACM international conference on Multimedia. Nara: Association for Computing Machinery, 2012: 973

[4] CHU Lingyang, JIANG Shuqiang, WANG Shuhui, et al. Robust spatial consistency graph model for partial duplicate image retrieval [J]. IEEE Transactions on Multimedia, 2013, 15(8): 1982. DOI: 10.1109/TMM.2013.2270455

[5] PARUI S, MITTAL A. Similarity-invariant sketch-based image retrieval in large databases [C]//Proceedings of the European Conference on Computer Vision. Zurich: Springer, 2014: 398. DOI: 10.1007/978-3-319-10599-4_26

[6]HE Ran, ZHANG Man, WANG Liang, et al. Cross-modal learning subspace learning via pairwise constraints[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5543. DOI: 10.1109/TIP.2015.2466106

[7]XU Zhongwen, YANG Yi, KASSIM A, et al. Cross-media relevance mining for evaluating text-based image search engine [C]// Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Chengdu: IEEE, 2014: 1. DOI: 10.1109/ICMEW.2014.6890606

[8]WU Dayan, LIU Jing, LI Bo, et al. Deep index-compatible hashing for fast image retrieval [C]// Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego: IEEE, 2018: 1. DOI: 10.1109/ICME.2018.8486463

[9]TANG Jinya, ZHANG Dongming, ZHANG Yongdong, et al. Region similarity arrangement for image retrieval[C]// Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle: IEEE, 2016: 1. DOI: 10.1109/ICME.2016.7552860

[10]DOU Zhi, WANG Ning, LI Baopu, et al. Dual color space guided sketch colorization[J]. IEEE Transactions on Image Processing. 2021, 30: 7292. DOI: 10.1109/TIP.2021.3104190

[11]YANG Lumin, ZHUANG Jiajie, FU Hongbo, et al. SketchGCN: Semantic sketch segmentation with graph convolutional networks[Z/OL]. [2020-03-02]. https://arxiv.org/pdf/2003.00678

[12]WANG Shengyu, BAU D, ZHU Junyan. Sketch your own GAN[Z/OL]. (2021-09-20). https://arxiv.org/abs/2108.02774

[13]KATO T, KURITA T, OTSU N, et al. A sketch retrieval method for full color image database-query by visual example [C]// Proceedings of the 11th IAPR International Conference on Pattern Recognition. Hague: IEEE, 1992: 530. DOI: 10.1109/ICPR.1992.201616

[14]PENG Yuxin, QI Jinwei. Show and tell in the loop: Cross-modal circular correlation learning[J]. IEEE Transactions on Multimedia, 2019, 21(6): 1538. DOI: 10.1109/TMM.2018.2877885

[15]RASTEGAR S, BAGHSHAH M S, RABIEE H R, et al. MDL-CW: A multimodal deep learning framework with cross weights [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2601. DOI: 10.1109/CVPR.2016.285

[16]DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005: 886. DOI: 10.1109/CVPR.2005.177

[17]EITZ M, HILDEBRAND K, BOUBEKEUR T, et al. Sketch-based image retrieval: Benchmark and bag-of-features descriptors [J]. IEEE Transactions on Visualization & Computer Graphics, 2011, 17(11): 1624. DOI: 10.1109/TVCG.2010.266.

[18]HU Rui, COLLOMOSSE J. A performance evaluation of gradient field HOG descriptor for sketch based image retrieval[J]. Computer Vision and Image Understanding, 2013, 117(7): 790. DOI: 10.1016/j.cviu.2013.02.005

[19]BUI T, COLLOMOSSE J. Scalable sketch-based image retrieval using color gradient features[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop. Santiago: IEEE, 2015: 1012. DOI: 10.1109/ICCVW.2015.133

[20]HU Rui, BARNARD M, COLLOMOSSE J. Gradient field descriptor for sketch based retrieval and localization [C]// Proceedings of the IEEE International Conference on Image Processing. Hong Kong: IEEE, 2010: 1025. DOI: 10.1109/ICIP.2010.5649331

[21]SAAVEDRA J M, BUSTOS B. An improved histogram of edge local orientations for sketch-based image retrieval[C]//Proceedings of the DAGM Conference on Pattern Recognition. Darmstadt: Springer, 2010: 432

[22]SAAVEDRA J M. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (S-HELO) [C]//Proceedings of the 2014 IEEE International Conference on Image Processing. Québec: IEEE, 2015: 2998. DOI: 10.1109/ICIP.2014.7025606

[23]SAAVEDRA J M, BARRIOS J M. Sketch based image retrieval using learned keyshapes (LKS)[C]//Proceedings of the British Machine Vision Conference. Swansea: BMVA Press, 2015: 1. DOI: 10.5244/C.29.164

[24]QI Yonggang, SONG Yizhe, XIANG Tao, et al. Making better use of edges via perceptual grouping[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1856. DOI: 10.1109/CVPR.2015.7298795

[25]LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C]//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2006: 2169. DOI: 10.1109/CVPR.2006.68

[26]ZHOU Yuan, HUO Shuwei, XIANG Wei, et al. Semi-supervised salient object detection using a linear feedback control system model [J]. IEEE Transactions on Cybernetics, 2019, 49(4): 1173. DOI: 10.1109/TCYB.2018.2793278

[27]HUO Shuwei, ZHOU Yuan, XIANG Wei, et al. Semi-supervised learning based on a novel iterative optimization model for saliency detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(1): 225. DOI: 10.1109/TNNLS.2018.2809702

[28]张博言, 钟勇. 一种基于多样性正实例的单目标跟踪算法[J]. 哈尔滨工业大学学报, 2020, 52(10): 135
ZHANG Boyan, ZHONG Yong. Single target tracking algorithm based on diverse positive instances[J]. Journal of Harbin Institute of Technology, 2020, 52(10): 135. DOI:10.11918/201907131

[29]EITZ M, HILDEBR K, BOUBEKEUR T, et al. A descriptor for large scale image retrieval based on sketched feature lines[C]// Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling. New Orleans: ACM Press, 2009: 29. DOI: 10.1145/1572741.1572747

[30]YU Qian, YANG Yongxin, SONG Yizhe, et al. Sketch-a-net that beats humans [C]//Proceedings of the British Machine Vision Conference. Swansea: BMVA Press, 2015: 1. DOI: 10.5244/C.29.7

[31]KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe: MIT Press, 2012: 1097

[32]QI Yonggang, SONG Yizhe, ZHANG Honggang, et al. Sketch-based image retrieval via Siamese convolutional neural network[C]// Proceedings of the 2016 IEEE International Conference on Image Processing. Phoenix: IEEE, 2016: 2460. DOI: 10.1109/ICIP.2016.7532801

[33]YU Qian, LIU Feng, SONG Yizhe, et al. Sketch me that shoe[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and

Pattern Recognition. Las Vegas: IEEE, 2016: 799. DOI: 10.1109/CVPR.2016.93

[34] BUI T, RIBEIRO L, PONTI M, et al. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network[J]. Computer Vision and Image Understanding, 2017, 164: 27. DOI: 10.1016/j.cviu.2017.06.007

[35] SANGKLOY P, BURNELL N, HAM C, et al. The sketchy database: Learning to retrieve badly drawn bunnies[J]. ACM Transactions on Graphics, 2016, 35(4): 1. DOI: 10.1145/2897824.2925954

[36] BUI T, RIBEIRO L S, PONTI M A, et al. Generalisation and sharing in triplet convnets for sketch based visual search[Z/OL]. [2016-11-16]. https://arxiv.org/pdf/1611.05301

[37] KUMAR M P, PACKER B, KOLLER D. Self-paced learning for latent variable models[C]// Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver: MIT Press, 2011: 1189

[38] BENGIO Y, LOURADOUR J, COLLOBERTR, et al. Curriculum learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. Québec: ACM Press, 2009: 41. DOI: 10.1145/1553374.1553380

[39] XU Dan, ALAMEDA-PINEDA X, SONG Jingkuan, et al. Cross-paced representation learning with partial curricula for sketch-based image retrieval[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4410. DOI: 10.1109/TIP.2018.2837381

[40] ZHOU Yuan, FENG Liyang, HOU Chunping, et al. Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing[J]. IEEE Transactions on Geoscience and Remote Sensing. 2017, 55(10): 5997. DOI: 10.1109/TGRS.2017.2718728

[41] LEI Pang, ZHU Shiai, NGO C W. Deep multimodal learning for affective analysis and retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 2008. DOI: 10.1109/TMM.2015.2482228

[42] ZHOU Yuan, XIE Xukai, KUNG S Y. Exploiting operation importance for differentiable neural architecture search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021: 1. DOI: 10.1109/TNNLS.2021.3072950

[43] ZHOU Yuan, YANG Jianxing, LI Hongru, et al. Adversarial learning for multiscale crowd counting under complex scenes[J]. IEEE Transactions on Cybernetics, 2020, 51(11): 5423. DOI: 10.1109/TCYB.2019.2956091

[44] BHATTACHARJEE S D, YUAN Junsong, HONG Weixiang, et al. Query adaptive instance search using object sketches[C]// Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam: ACM Press, 2016:1306. DOI: 10.1145/2964284.2964317

[45] LEORDEANU M, SUKTHANKAR R, SMINCHISESCU C. Efficient closed-form solution to generalized boundary detection[C]// Proceedings of the European Conference on Computer Vision. Florence: Springer, 2012: 516

[46] EITZ M, HAYS J, ALEXA M. How do humans sketch objects[J]. ACM Transactions of Graphics, 2012, 31(4): 1. DOI: 10.1145/2185520.2185540

[47] TOLIAS G, CHUM O. Asymmetric feature maps with application to sketch based retrieval[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2377. DOI: 10.1109/CVPR.2017.655

[48] BHATTACHARJEE S D, YUAN Junsong, HUANG Yicheng, et al. Query adaptive multiview object instance search and localization using sketches[J]. IEEE Transactions on Multimedia, 2018, 20(10): 2761. DOI: 10.1109/TMM.2018.2814338

[49] SEDDATI O, DUPONT S, MAHMOUDI S. Quadruplet networks for sketch-based image retrieval[C]// Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. Chengdu: ACM Press, 2017: 184. DOI:10.1145/3078971.3078985

[50] BUI T, RIBEIRO L, PONTI M, et al. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression[J]. Computers & Graphics, 2018, 71: 77. DOI: 10.1016/j.cag.2017.12.006

[51] ZHOU Yuan, DU Xiaoting, WANG Mingfei, et al. Cross-scale residual network: a general framework for image super-resolution, denoising, and deblocking[J]. IEEE Transactions on Cybernetics, 2020: 1. DOI:10.1109/TCYB.2020.3044374

[52] JOLY A, BUISSON O. Logo retrieval with acontrario visual query expansion[C]//Proceedings of the 17th ACM International Conference on Multimedia. Vancouver: Springer, 2009: 581. DOI: 10.1145/1631272.1631361

（编辑　苗秀芝）